第三章

非球形扰动与广义最小二乘

Outline

1. 广义最小二乘
2. Feasible Generalized Least Squares
3. 条件异方差检验
4. 自相关检验
5. Heteroskedasticity and Autocorrelation Consistent (HAC) Covariance Matrix Estimator
6. Clustered Sampling

广义最小二乘

Generalized Least Squares (GLS)

- Assumption 1.4 states that the $n \times n$ matrix of conditional second moments $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})$ is spherical, that is, proportional to the identity matrix.
- If the error is not (conditionally) homoskedastic, the values of the diagonal elements of $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})$ are not the same.
- If there is correlation in the error term between observations, the values of the off-diagonal elements are not zero.
- The model that results when Assumption 1.4 is replaced by

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\mathbf{V}(\mathbf{X}), \quad \mathbf{V}(\mathbf{X}) \text{ nonsingular and know,} \tag{1}$$

is called the generalized regression model.

Consequence of Relaxing Assumption 1.4

1. The Gauss-Markov Theorem no longer holds for the OLS estimator:

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{b}|\mathbf{X}) &= \mathrm{Var}(\boldsymbol{b}-\boldsymbol{\beta}|\mathbf{X}) \\
&= \mathrm{Var}(\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}), \quad \mathbf{A}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{A}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})\mathbf{A}' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

2. The t-ratio is not distributed as the t distribution. Thus, the t-test is no longer valid. The same comments apply to the F-test.
3. The OLS estimator is still unbiased (small sample) or consistent (large sample). For example,

$$\mathbb{E}(\boldsymbol{b}-\boldsymbol{\beta}|\mathbf{X}) = \mathbb{E}(\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{A}\mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}.$$

Efficient Estimation with Known $\mathbf{V}$

Let the characteristic roots and vectors of $\mathbf{V}$ be arranged in a diagonal matrix $\mathbf{\Lambda}$ and an orthogonal matrix $\mathbf{C}$. Since $\mathbf{V}$ is symmetric and positive definite, then

$$\mathbf{V} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}'.$$

Therefore,

$$\mathbf{V}^{-1} = \mathbf{P}'\mathbf{P},$$

where $\mathbf{P} \equiv \mathbf{C}\mathbf{\Lambda}^{-1/2}$.

Now consider creating a new regression model by transforming $(\mathbf{y}, \mathbf{X}, \boldsymbol{\varepsilon})$ by $\mathbf{P}$ as

$$\tilde{\mathbf{y}} \equiv \mathbf{P}\mathbf{y}, \quad \tilde{\mathbf{X}} \equiv \mathbf{P}\mathbf{X}, \quad \tilde{\boldsymbol{\varepsilon}} \equiv \mathbf{P}\boldsymbol{\varepsilon}.$$

- Assumption 1.1 for $(\mathbf{y}, \mathbf{X}, \boldsymbol{\varepsilon})$ implies that $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\boldsymbol{\varepsilon}})$ too satisfies linearity:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}. \tag{2}$$

- Strict exogeneity is satisfied because

$$\mathbb{E}(\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\mathbf{X}}) = \mathbb{E}(\tilde{\boldsymbol{\varepsilon}} \mid \mathbf{X}) = \mathbf{P}\mathbb{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}.$$

- Because $\mathbf{V}$ is positive definite, the no-multicollinearity assumption is also satisfied.
- Assumption 1.4 is satisfied for the transformed model because

$$\begin{aligned} \mathbb{E}(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}' \mid \tilde{\mathbf{X}}) &= \mathbb{E}(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}' \mid \mathbf{X}) = \mathbf{P}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X})\mathbf{P}' \\ &= \sigma^2\mathbf{P}\mathbf{V}\mathbf{P}' = \sigma^2\mathbf{I}_n. \end{aligned}$$

- $\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\mathbf{X}}$ is normal if $\boldsymbol{\varepsilon} \mid \mathbf{X}$ is normal.

The Gauss-Markov Theorem for the transformed model implies that the BLUE of $\boldsymbol{\beta}$ for the generalized regression model is the OLS estimator applied to (2):

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

This is the GLS estimator.

Theorem (finite-sample properties of GLS)

(a) *(Unbiasedness) Under Assumptions 1.1-1.3,* $\mathbb{E}(\hat{\boldsymbol{\beta}}_{\text{GLS}}|\mathbf{X}) = \boldsymbol{\beta}$.

(b) *(Expression for the variance) Under Assumptions 1.1-1.3 and the assumption (1),* $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}|\mathbf{X}) = \sigma^2 \cdot (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$.

(c) *(Efficiency of GLS) Under the same set of assumptions as in (b), the GLS estimator is efficient in that the conditional variance of any unbiased estimator that is linear in* $\mathbf{y}$ *is greater than or equal to* $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}|\mathbf{X})$ *in the matrix sense.*

Proof.

(a) Note that $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}$. Therefore,

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \boldsymbol{\beta}|\mathbf{X}) = \mathbb{E}\left[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}|\mathbf{X}\right] = \mathbf{0}.$$

(b)

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}|\mathbf{X}) &= \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \boldsymbol{\beta}|\mathbf{X}) \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&= \sigma^2 \cdot (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.
\end{aligned}$$

(c) Efficiency of GLS is implied by the fact that the transformed model satisfies Assumptions 1.1-1.4. □

A Special Case: Weighted Least Squares (WLS)

When there is no correlation in the error term between observations, $\mathbf{V}$ is diagonal. Let $v_i(\mathbf{X})$ be the i-th diagonal element of $\mathbf{V}(\mathbf{X})$. So

$$\mathbb{E}(\varepsilon_i^2 \mid \mathbf{X}) = \sigma^2 \cdot v_i(\mathbf{X}).$$

It is easy to see that $\mathbf{P}$ is also diagonal, with $1/\sqrt{v_i(\mathbf{X})}$ in the i-th diagonal. Thus, $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ is given by

$$\tilde{y}_i = \frac{y_i}{\sqrt{v_i(\mathbf{X})}}, \quad \tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\sqrt{v_i(\mathbf{X})}} \quad (i = 1, 2, \ldots, n).$$

Therefore, efficient estimation under a known form of heteroskedasticity is first to weight each observation by the reciprocal of the square root of the variance $v_i(\mathbf{X})$ and then apply OLS. This is called the weighted regression (or the weighted least squares (WLS)).

An important further special case is the case of a random sample where $\{y_i, \mathbf{x}_i\}$ is i.i.d. across i. In this case,

- the error is unconditionally homoskedastic
- the error can be conditionally heteroskedastic
- $\mathbb{E}(\varepsilon_i^2 \mid \mathbf{X})$ depends only on $\mathbf{x}_i$, and the functional form of $E(\varepsilon_i^2 \mid \mathbf{x}_i)$ is the same across i.

Thus

$$v_i(\mathbf{X}) = v(\mathbf{x}_i).$$

So the knowledge of $\mathbf{V}(\cdot)$ comes down to a single function of K variables, $v(\cdot)$.

Remark

1. The finite-sample properties of GLS rest on the assumption that the regressors in the generalized regression model are strictly exogenous.
2. If the regressors are not strictly exogenous but are merely predetermined, the GLS procedure to correct for serial correlation can make the estimator inconsistent.
3. If we do not know the function $\mathbf{V}(\mathbf{X})$, we can estimate its functional form from the sample. This approach is called the Feasible Generalized Least Squares (FGLS). Very little is known about the finite-sample properties of the FGLS estimator.
4. For FGLS, with an unrestricted matrix $\mathbf{V}(\mathbf{X})$, there are $n(n+1)/2$ additional parameters to estimate. This number is far too many to estimate with n observations. Some structure must be imposed on the model.

Feasible Generalized Least Squares

To simplify the discussion, we assume that $\{y_i, \mathbf{x}_i\}$ is i.i.d., and

$$\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i) = \mathbf{z}_i'\boldsymbol{\alpha}, \tag{3}$$

where $\mathbf{z}_i$ is a function of $\mathbf{x}_i$.

WLS with known $\boldsymbol{\alpha}$: The WLS estimator is given by

$$\hat{\boldsymbol{\beta}}(\mathbf{V}) \equiv \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{x}}_i\tilde{y}_i = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

with

$$\mathbf{V} \equiv \begin{bmatrix} \mathbf{z}_1'\boldsymbol{\alpha} & & \\ & \ddots & \\ & & \mathbf{z}_n'\boldsymbol{\alpha} \end{bmatrix}.$$

If Assumption 2.3 is strengthened by the condition that

$$\mathbb{E}(\varepsilon_i \mid \mathbf{x}_i) = 0,$$

then $\mathbb{E}(\tilde{\varepsilon}_i \mid \tilde{\mathbf{x}}_i) = 0$. Therefore, provided that $\mathbb{E}(\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i')$ is nonsingular, the WLS estimator is consistent and asymptotically normal (Assumptions 2.1-2.5 are satisfied), and the asymptotic variance is

$$\begin{aligned}
\operatorname{Avar}(\hat{\boldsymbol{\beta}}(\mathbf{V})) &= \mathbb{E}(\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i')^{-1} \\
&= \operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i'\right)^{-1} \\
&= \operatorname{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}.
\end{aligned}$$

Regression of e_i^2 on $\mathbf{z}_i$ provides a consistent estimate of $\boldsymbol{\alpha}$: If $\boldsymbol{\alpha}$ is unknown, it can be estimated by running a separate regression.

- Note that (3) can be written as a regression equation:

$$\varepsilon_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i, \tag{4}$$

  where $\eta_i \equiv \varepsilon_i^2 - \mathbb{E}(\varepsilon_i^2|\mathbf{x}_i)$. By construction, $\mathbb{E}(\eta_i|\mathbf{x}_i) = 0$, which implies that $\mathbb{E}(\eta_i|\mathbf{z}_i) = 0$.

- Hence, provided that $\mathbb{E}(\mathbf{z}_i\mathbf{z}_i')$ is nonsingular, the OLS estimator of $\boldsymbol{\alpha}$ is consistent and asymptotically normal.
- When ε_i is replaced by e_i in the regression (4), the OLS estimator, call it $\hat{\boldsymbol{\alpha}}$, is consistent for $\boldsymbol{\alpha}$.

WLS with Estimated $\boldsymbol{\alpha}$

Step 1: Estimate the equation $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ by OLS and compute the OLS residuals e_i.

Step 2: Regress e_i^2 on $\mathbf{z}_i$, to obtain the OLS coefficient estimate $\hat{\boldsymbol{\alpha}}$.

Step 3: Re-estimate the equation $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ by WLS, using $1/\sqrt{\mathbf{z}_i'\hat{\boldsymbol{\alpha}}}$ as the weight for observation i.

It can be shown that the WLS estimator is asymptotically more efficient than the OLS estimator.

The superiority of WLS over OLS, however, rests on the premise that the sample size is sufficiently large and the functional form of the conditional second moment is correctly specified.

条件异方差检验

White Test

Recall that $\hat{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i'$ is consistent for $\mathbf{S} = \mathbb{E}(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i')$. Under conditional homoskedasticity, $s^2 \mathbf{S}_{\mathbf{xx}}$ is consistent for $\mathbf{S}$. So the difference between the two consistent estimators should vanish:

$$\hat{\mathbf{S}} - s^2 \mathbf{S}_{\mathbf{xx}} = \frac{1}{n}\sum_{i=1}^{n}(e_i^2 - s^2)\mathbf{x}_i \mathbf{x}_i' \underset{p}{\to} \mathbf{0}.$$

Let $\boldsymbol{\psi}_i$ be a vector collecting unique and nonconstant elements of the $K \times K$ symmetric matrix $\mathbf{x}_i \mathbf{x}_i'$. Then,

$$\mathbf{c}_n \equiv \frac{1}{n}\sum_{i=1}^{n}(e_i^2 - s^2)\boldsymbol{\psi}_i \underset{p}{\to} \mathbf{0}.$$

Under some conditions appropriate for a CLT to be applicable, $\sqrt{n}\mathbf{c}_n$ converges in distribution to a normal distribution with mean zero and some asymptotic variance $\mathbf{B}$, so for any consistent estimator $\hat{\mathbf{B}}$ of $\mathbf{B}$,

$$n \cdot \mathbf{c}_n' \hat{\mathbf{B}}^{-1} \mathbf{c}_n \underset{d}{\to} \chi^2(m),$$

where m is the dimension of $\mathbf{c}_n$.

For a certain choice of $\hat{\mathbf{B}}$, this statistic can be computed as nR^2 from the following auxiliary regression:

$$\text{regress } e_i^2 \text{ on a constant and } \boldsymbol{\psi}_i. \tag{5}$$

Theorem (White's Test for Conditional Heteroskedasticity)

In addition to Assumptions 2.1 and 2.4, suppose that (a) $\{y_i, \mathbf{x}_i\}$ is i.i.d. with finite $\mathbb{E}(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i')$, (b) ε_i is independent of $\mathbf{x}_i$, and (c) a certain condition holds on the moments of ε_i and $\mathbf{x}_i$. Then,

$$nR^2 \underset{d}{\to} \chi^2(m),$$

where R^2 is the R^2 from the auxiliary regression (5), and m is the dimension of $\boldsymbol{\psi}_i$.

Breusch-Pagan Test

Consider the linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{6}$$

where ε_i is normally and independently distributed with mean zero and variance

$$\sigma_i^2 = \sigma^2 \cdot h(\mathbf{z}_i'\boldsymbol{\alpha}),$$

where $\boldsymbol{\alpha}$ is $p \times 1$ vector and the first element is 1. The null hypothesis of homoscedasticity can be written as

$$\mathbb{H}_0 : \alpha_2 = \cdots = \alpha_p = 0.$$

It is also assumed that $\mathbf{x}_i$ and $\mathbf{z}_i$ are exogenous.

Breusch-Pagan Test

Step 1: Run regression (6), obtain residuals e_i's.

Step 2: Obtain g_i which is given by

$$g_i \equiv \frac{e_i^2}{s^2}, \quad \text{with } s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n}$$

for $i = 1, \ldots, n$.

Step 3: Regress g_i on $\mathbf{z}_i$, obtain the explained sum of squares which is $\boldsymbol{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{g} - n$, where $\boldsymbol{g} = (g_1, \ldots, g_n)'$, $\mathbf{Z} = (\mathbf{z}_1', \ldots, \boldsymbol{z}_n')'$.

Step 4: Construct the test statistic, under $\mathbb{H}_0$,

$$\text{LM} \equiv \frac{1}{2}\left[\boldsymbol{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{g} - n\right] \underset{d}{\to} \chi^2(p-1).$$

自相关检验

- When the regressors include a constant (true in virtually all known applications), Assumption 2.5 implies that the error term is a scalar martingale difference sequence.
- If the error is found to be serially correlated, that is an indication of a failure of Assumption 2.5.
- Serial correlation has traditionally been an important subject in econometrics, and there are available a number of tests for serial correlation.
- However, some of them (e.g., Durbin-Watson test) require that the regressors be strictly exogenous.

Digression: Box-Pierce and Ljung-Box

Suppose we have a sample $\{z_1, \ldots, z_n\}$ drawn from a scalar covariance-stationary process. The sample j-th order autocovariance is

$$\hat{\gamma}_j \equiv \frac{1}{n} \sum_{t=j+1}^{n} (z_t - \bar{z}_n)(z_{t-j} - \bar{z}_n),$$

where

$$\bar{z}_n \equiv \frac{1}{n} \sum_{t=1}^{n} z_t.$$

The sample j-th order autocorrelation coefficient, $\hat{\rho}_j$, is defined as

$$\hat{\rho}_j \equiv \frac{\hat{\gamma}_j}{\hat{\gamma}_0}.$$

If $\{z_t\}$ is ergodic stationary, then it is easy to show that $\hat{\gamma}_j$ is consistent for γ_j, and $\hat{\rho}_j$ is consistent for ρ_j.

Theorem

Suppose $\{z_t\}$ *can be written as* $\mu + \varepsilon_t$, *where* ε_t *is a stationary martingale difference sequence with "own" conditional homoskedasticity:*

$$\textit{(own conditional homoskedasticity)}\ \ \mathbb{E}(\varepsilon_t^2 \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) = \sigma^2 > 0.$$

Then,

$$\sqrt{n}\hat{\boldsymbol{\gamma}} \underset{d}{\to} N(\mathbf{0}, \sigma^4 \mathbf{I}_p) \quad \textit{and} \quad \sqrt{n}\hat{\boldsymbol{\rho}} \underset{d}{\to} N(\mathbf{0}, \mathbf{I}_p),$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$ *and* $\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_p)'$. *(Cf. [Hayashi, 2000], p.171, Q9)*

For testing $\mathbb{H}_0 : \rho_1 = \cdots = \rho_p = 0$, we have:

- Box-Pierce Q-test:

$$Q \equiv n \sum_{j=1}^{p} \hat{\rho}_j^2 = \sum_{j=1}^{p} (\sqrt{n}\hat{\rho}_j)^2 \underset{d}{\to} \chi^2(p).$$

- Ljung-Box Q-test:

$$Q^* \equiv n(n+2) \sum_{j=1}^{p} \frac{\hat{\rho}_j^2}{n-j} = \sum_{j=1}^{p} \frac{n+2}{n-j} (\sqrt{n}\hat{\rho}_j)^2 \underset{d}{\to} \chi^2(p).$$

Sample Autocorrelations Calculated from Residuals

For the regression model described by Assumptions 2.1-2.5, if ε_t were observable, we would calculate

$$\tilde{\rho}_j \equiv \frac{\tilde{\gamma}_j}{\tilde{\gamma}_0} \quad (j = 1, 2, \ldots),$$

where

$$\tilde{\gamma}_j \equiv \frac{1}{n} \sum_{t=j+1}^{n} \varepsilon_t \varepsilon_{t-j} \quad (j = 0, 1, 2, \ldots).$$

Consider the realistic case where we do not observe ε_t. We can replace ε_t by e_t and calculate

$$\hat{\rho}_j \equiv \frac{\hat{\gamma}_j}{\hat{\gamma}_0} \quad (j = 1, 2, \ldots),$$

where

$$\hat{\gamma}_j \equiv \frac{1}{n} \sum_{t=j+1}^{n} e_t e_{t-j} \quad (j = 0, 1, 2, \ldots).$$

Is it all right to use residual-based Q test derived from $\{\hat{\rho}_j\}$ for testing serial correlation?

Answer: YES! But only if the regressors are strictly exogenous.

To see this, recall that

$$e_t = \varepsilon_t - \mathbf{x}_t'(\boldsymbol{b} - \boldsymbol{\beta}).$$

Therefore,

$$\begin{aligned}
\hat{\gamma}_j &\equiv \frac{1}{n}\sum_{t=j+1}^{n} e_t e_{t-j} = \frac{1}{n}\sum_{t=j+1}^{n} [\varepsilon_t - \mathbf{x}_t'(\boldsymbol{b} - \boldsymbol{\beta})][\varepsilon_{t-j} - \mathbf{x}_{t-j}'(\boldsymbol{b} - \boldsymbol{\beta})] \\
&= \tilde{\gamma}_j - \frac{1}{n}\sum_{t=j+1}^{n} (\mathbf{x}_{t-j}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-j})'(\boldsymbol{b} - \boldsymbol{\beta}) \qquad (7) \\
&\quad + (\boldsymbol{b} - \boldsymbol{\beta})' \left(\frac{1}{n}\sum_{t=j+1}^{n} \mathbf{x}_t\mathbf{x}_{t-j}'\right)(\boldsymbol{b} - \boldsymbol{\beta}).
\end{aligned}$$

Since $\boldsymbol{b} - \boldsymbol{\beta} \underset{p}{\to} \mathbf{0}$, $\hat{\gamma}_j - \tilde{\gamma}_j \underset{p}{\to} 0$ $(j = 0, 1, 2, \ldots)$, and the difference between $\tilde{\rho}_j$ and $\hat{\rho}_j$ vanishes in large samples.

However, by multiplying both sides of (7) by $\sqrt{n}$, we have

$$\begin{aligned}\sqrt{n}\hat{\gamma}_j &= \sqrt{n}\tilde{\gamma}_j - \frac{1}{n}\sum_{t=j+1}^{n}(\mathbf{x}_{t-j}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-j})'\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta}) \\ &\quad + \underbrace{\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})'\left(\frac{1}{n}\sum_{t=j+1}^{n}\mathbf{x}_t\mathbf{x}'_{t-j}\right)(\boldsymbol{b}-\boldsymbol{\beta})}_{\xrightarrow[p]{} 0}.\end{aligned}$$

Regarding the second term, we have

$$\frac{1}{n}\sum_{t=j+1}^{n}(\mathbf{x}_{t-j}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-j}) \xrightarrow[p]{} \mathbb{E}(\mathbf{x}_{t-j}\varepsilon_t) + \mathbb{E}(\mathbf{x}_t\varepsilon_{t-j}).$$

If the regressors are strictly exogenous in the sense that $\mathbb{E}(\mathbf{x}_t\varepsilon_s) = \mathbf{0}$ for all t, s, then

$$\mathbb{E}(\mathbf{x}_{t-j}\varepsilon_t) + \mathbb{E}(\mathbf{x}_t\varepsilon_{t-j}) = \mathbf{0}.$$

So the difference between $\sqrt{n}\tilde{\rho}_j$ and $\sqrt{n}\hat{\rho}_j$ vanishes, which means that the Q statistic calculated from $\{e_t\}$ is asymptotically chi-squared, and we can use this residual-based Q to test for serial correlation.

Testing with Predetermined, but Not Strictly Exogenous, Regressors

When the regressors are not strictly exogenous, we need to modify the Q statistic to restore its asymptotic distribution. Consider two restrictions:

Stronger form of predeterminedness:

$$\mathbb{E}(\varepsilon_t \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \mathbf{x}_t, \mathbf{x}_{t-1}, \ldots) = 0. \tag{8}$$

Stronger form of conditional homoskedasticity:

$$\mathbb{E}(\varepsilon_t^2 \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \mathbf{x}_t, \mathbf{x}_{t-1}, \ldots) = \sigma^2 > 0. \tag{9}$$

Remark:

- (8) is stronger than Assumption 2.3 and implies that $\mathbf{g}_t$ is an m.d.s.
- (9) is obviously stronger than Assumption 2.7.

Theorem (testing for serial correlation with predetermined regressors)

Suppose that Assumptions 2.1, 2.2, 2.4, (8), and (9) are satisfied. Then,

$$\sqrt{n}\hat{\boldsymbol{\gamma}} \underset{d}{\to} N(\mathbf{0}, \sigma^4 \cdot (\mathbf{I}_p - \boldsymbol{\Phi})) \quad \textit{and} \quad \sqrt{n}\hat{\boldsymbol{\rho}} \underset{d}{\to} N(\mathbf{0}, \mathbf{I}_p - \boldsymbol{\Phi}),$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$, $\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_p)'$, *and* ϕ_{jk}, *the* (j,k)*-th element of the* $p \times p$ *matrix* $\boldsymbol{\Phi}$, *is given by*

$$\phi_{jk} = \mathbb{E}(\mathbf{x}_t \varepsilon_{t-j})' \mathbb{E}(\mathbf{x}_t \mathbf{x}_t')^{-1} \mathbb{E}(\mathbf{x}_t \varepsilon_{t-k}) / \sigma^2.$$

By the Ergodic Theorem, matrix $\boldsymbol{\Phi}$ is consistently estimated by

$$\hat{\boldsymbol{\Phi}} \equiv (\hat{\phi}_{jk}), \quad \hat{\phi}_{jk} \equiv \bar{\boldsymbol{\mu}}_j' \mathbf{S}_{\mathbf{xx}}^{-1} \bar{\boldsymbol{\mu}}_k / s^2 \quad (j,k = 1, 2, \ldots, p),$$

where

$$s^2 \equiv \frac{1}{n-K} \sum_{t=1}^{n} e_t^2, \quad \bar{\boldsymbol{\mu}}_j \equiv \frac{1}{n} \sum_{t=j+1}^{n} \mathbf{x}_t \cdot e_{t-j}.$$

It follows that

$$\text{modified Box-Pierce } Q \equiv n \cdot \hat{\boldsymbol{\rho}}' (\mathbf{I}_p - \hat{\boldsymbol{\Phi}})^{-1} \hat{\boldsymbol{\rho}} \underset{d}{\to} \chi^2(p).$$

Proof.

Note that

$$
\begin{aligned}
\sqrt{n}\hat{\gamma}_j &= \sqrt{n}\tilde{\gamma}_j - \frac{1}{n}\sum_{t=j+1}^{n}(\mathbf{x}_{t-j}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-j})'\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta}) \\
&\quad + \sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})'\left(\frac{1}{n}\sum_{t=j+1}^{n}\mathbf{x}_t\mathbf{x}_{t-j}'\right)(\boldsymbol{b}-\boldsymbol{\beta}) \\
&\underset{a}{\sim} \sqrt{n}\tilde{\gamma}_j - \frac{1}{n}\sum_{t=j+1}^{n}(\mathbf{x}_{t-j}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-j})'\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta}) \\
&\underset{a}{\sim} \sqrt{n}\tilde{\gamma}_j - \mathbb{E}(\mathbf{x}_{t-j}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-j})'\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta}) \\
&= \sqrt{n}\tilde{\gamma}_j - \boldsymbol{\mu}_j'\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta}) \quad \text{where } \boldsymbol{\mu}_j \equiv \mathbb{E}(\mathbf{x}_t\varepsilon_{t-j}) \\
&= \sqrt{n}\tilde{\gamma}_j - \boldsymbol{\mu}_j'\sqrt{n}\mathbf{S}_{\mathbf{xx}}^{-1}\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t\varepsilon_t \\
&= \sqrt{n}\frac{1}{n}\sum_{t=j+1}^{n}\varepsilon_t\varepsilon_{t-j} - \boldsymbol{\mu}_j'\sqrt{n}\mathbf{S}_{\mathbf{xx}}^{-1}\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t\varepsilon_t \\
&\underset{a}{\sim} \sqrt{n}\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t\varepsilon_{t-j} - \boldsymbol{\mu}_j'\sqrt{n}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t\varepsilon_t = \mathbf{c}_j'\sqrt{n}\bar{\mathbf{g}}_j,
\end{aligned}
$$

Proof.

where

$$\boldsymbol{c}_j \equiv \begin{bmatrix} 1 \\ -\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_j \end{bmatrix}, \quad \bar{\mathbf{g}}_j \equiv \frac{1}{n}\sum_{t=1}^{n}\mathbf{g}_{jt}, \quad \mathbf{g}_{jt} \equiv \begin{bmatrix} \varepsilon_{t-j}\varepsilon_t \\ \mathbf{x}_t\varepsilon_t \end{bmatrix}.$$

Thus, letting $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$, we have proved that

$$\sqrt{n}\hat{\boldsymbol{\gamma}} \underset{a}{\sim} \mathbf{C}'\sqrt{n}\bar{\mathbf{g}}, \tag{10}$$

where

$$\mathbf{C} \equiv \begin{bmatrix} \mathbf{c}_1 & & \\ & \ddots & \\ & & \mathbf{c}_p \end{bmatrix}, \quad \bar{\mathbf{g}} \equiv \frac{1}{n}\sum_{t=1}^{n}\mathbf{g}_t, \quad \mathbf{g}_t \equiv \begin{bmatrix} \mathbf{g}_{1t} \\ \vdots \\ \mathbf{g}_{pt} \end{bmatrix}.$$

Using LIE and (8), we can show that $\mathbf{g}_t$ is a m.d.s. Clearly, $\mathbf{g}_t$ is also ergodic stationary. Therefore,

$$\sqrt{n}\bar{\mathbf{g}} \underset{d}{\to} N(\mathbf{0}, \mathbb{E}(\mathbf{g}_t\mathbf{g}_t')). \tag{11}$$

The next step is to calculate $\mathbb{E}(\mathbf{g}_t\mathbf{g}_t')$. Its (j,k) block is

$$\mathbb{E}(\mathbf{g}_{jt}\mathbf{g}_{kt}') = \begin{bmatrix} \mathbb{E}(\varepsilon_t^2\varepsilon_{t-j}\varepsilon_{t-k}) & \mathbb{E}(\mathbf{x}_t'\varepsilon_t^2\varepsilon_{t-j}) \\ \mathbb{E}(\mathbf{x}_t\varepsilon_t^2\varepsilon_{t-k}) & \mathbb{E}(\varepsilon_t^2\mathbf{x}_t\mathbf{x}_t') \end{bmatrix}.$$

Proof.

By using LIE, (8), and (9), it is easy to show that

$$\mathbb{E}(\mathbf{g}_{jt}\mathbf{g}_{kt}') = \begin{bmatrix} \sigma^4\delta_{jk} & \sigma^2\boldsymbol{\mu}_j' \\ \sigma^2\boldsymbol{\mu}_k & \sigma^2\boldsymbol{\Sigma}_{\mathbf{xx}} \end{bmatrix},$$

where δ_{jk} is 1 if $j = k$ and 0 otherwise. From (10) and (11), we have

$$\text{Avar}(\hat{\boldsymbol{\gamma}}) = \mathbf{C}'\mathbb{E}(\mathbf{g}_t\mathbf{g}_t')\mathbf{C}.$$

Its (j, k) element is

$$\mathbf{c}_j'\mathbb{E}(\mathbf{g}_{jt}\mathbf{g}_{kt}')\mathbf{c}_k = \sigma^4 \cdot [\delta_{jk} - \boldsymbol{\mu}_j'\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_k/\sigma^2].$$

So

$$\sqrt{n}\hat{\boldsymbol{\gamma}} \underset{d}{\to} N(\mathbf{0}, \sigma^4 \cdot (\mathbf{I}_p - \boldsymbol{\Phi})).$$

Since $\sqrt{n}\hat{\boldsymbol{\rho}} \underset{a}{\sim} \sqrt{n}\hat{\boldsymbol{\gamma}}/\sigma^2$, the limiting distribution of $\sqrt{n}\hat{\boldsymbol{\rho}}$ is the same as that of $\sqrt{n}\hat{\boldsymbol{\gamma}}/\sigma^2$, which is $N(\mathbf{0}, \mathbf{I}_p - \boldsymbol{\Phi})$. □

An Auxiliary Regression-Based Test

Consider the following auxiliary regression:

$$\text{regress } e_t \text{ on } \mathbf{x}_t, e_{t-1}, e_{t-2}, \ldots, e_{t-p}. \tag{12}$$

From this regression, we can calculate the F statistic for the hypothesis that the p coefficients of $e_{t-1}, \ldots, e_{t-p}$ are all zero.

- Under the hypothesis of the previous theorem, the modified Q statistic is asymptotically equivalent to $p \cdot F$, so $p \cdot F$, too, is asymptotically chi-squared. (Cf. [Hayashi, 2000], p.173, Q11)
- This $p \cdot F$ statistic, in turn, is asymptotically equivalent to nR^2 from the auxiliary regression. The test based on nR^2 is called the Breusch-Godfrey test for serial correlation.

We now prove the last statement.

Proof.

Recall that

$$p \cdot F = \frac{SSR_R - SSR_U}{SSR_U/(n-K-p)} = (n-K-p)\frac{SSR_R - SSR_U}{SSR_U},$$

where SSR_U is from (12) and SSR_R is from the following restricted regression:

$$\text{regress } e_t \text{ on } \mathbf{x}_t. \tag{13}$$

However, since e_t is the OLS residual, $\mathbf{x}_t$ in (13) has no explanatory power. So $SSR_R = \boldsymbol{e}'\boldsymbol{e}$, and $p \cdot F$ is numerically identical to

$$(n-K-p)\frac{R^2_{uc}}{1-R^2_{uc}},$$

where R^2_{uc} is the uncentered R^2 for (12).

Proof.

Since the sample mean of e_t is zero, we also have

$$p \cdot F = (n - K - p)\frac{R^2}{1 - R^2},$$

where R^2 is equal to the R^2 for (12). Solving this equation for R^2 and multiplying both sides by n, we obtain

$$nR^2 = \frac{n}{n - K - p} \cdot \frac{1}{1 + \dfrac{p \cdot F}{n - K - p}} \cdot p \cdot F.$$

It is easy to see that $p \cdot F - nR^2 \underset{p}{\to} 0$. Therefore, nR^2 from (12) is asymptotically $\chi^2(p)$. □

Heteroskedasticity and Autocorrelation Consistent (HAC) Covariance Matrix Estimator

Asymptotics for Sample Means of Serially Correlated Processes

Consider the sample mean $\bar{y} \equiv \frac{1}{n}\sum_{t=1}^{n} y_t$ for serially correlated process $\{y_t\}$.

Theorem (LLN for covariance-stationary processes with vanishing autocovariances)

Let $\{y_t\}$ be covariance-stationary with mean μ and $\{\gamma_j\}$ be the autocovariances of $\{y_t\}$. Then

(a) $\bar{y} \underset{m.s.}{\to} \mu$ *as* $n \to \infty$, *if* $\lim_{j\to\infty} \gamma_j = 0$.

(b) $\lim_{n\to\infty} \mathrm{Var}(\sqrt{n}\bar{y}) = \sum_{j=-\infty}^{\infty} \gamma_j < \infty$, *if* $\{\gamma_j\}$ *is summable.*

For a covariance-stationary process $\{y_t\}$, define the long-run variance to be the limit as $n \to \infty$ of $\mathrm{Var}(\sqrt{n}\bar{y})$.

Note that

$$\begin{aligned}\mathrm{Var}(\sqrt{n}\bar{y}) &= \frac{1}{n}\mathrm{Var}\left(\sum_{t=1}^{n} y_t\right) = \frac{1}{n}\left(n\gamma_0 + \sum_{t=1}^{n}\sum_{s\neq t}^{n}\mathrm{Cov}(y_t, y_s)\right) \\ &= \gamma_0 + \frac{2}{n}\sum_{t=1}^{n}\sum_{s<t}^{n}\mathrm{Cov}(y_t, y_s) = \gamma_0 + \frac{2}{n}\sum_{j=1}^{n-1}(n-j)\gamma_j \\ &= \gamma_0 + 2\sum_{j=1}^{n-1}(1-\frac{j}{n})\gamma_j.\end{aligned}$$

Theorem (CLT for MA(∞))

Let $y_t = \mu + \sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$ *where* $\{\varepsilon_t\}$ *is independent white noise and* $\sum_{j=0}^{\infty}|\psi_j| < \infty$. *Then*

$$\sqrt{n}(\bar{y}-\mu) \underset{d}{\to} N\left(0, \sum_{j=-\infty}^{\infty}\gamma_j\right).$$

Theorem

Let $\bar{\mathbf{y}} \equiv \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t$ *be the sample mean of a vector process* $\{\mathbf{y}_t\}$.

(a) $\bar{\mathbf{y}} \underset{m.s.}{\to} \boldsymbol{\mu}$ *if each diagonal element of* $\boldsymbol{\Gamma}_j$ *goes to zero as* $j \to \infty$.

(b) $\lim_{n\to\infty} \mathrm{Var}(\sqrt{n}\bar{\mathbf{y}}) = \sum_{j=-\infty}^{\infty}\boldsymbol{\Gamma}_j$, *if* $\{\boldsymbol{\Gamma}_j\}$ *is summable.*

(c) *If* $\{\mathbf{y}_t\}$ *is vector MA*(∞) *with absolutely summable coefficients and* $\{\boldsymbol{\varepsilon}_t\}$ *is vector independent white noise, then*

$$\sqrt{n}(\bar{\mathbf{y}} - \boldsymbol{\mu}) \underset{d}{\to} N\left(\mathbf{0}, \sum_{j=-\infty}^{\infty}\boldsymbol{\Gamma}_j\right).$$

HAC Covariance Matrix Estimator

- Recall that in Assumption 2.5, we assume that $\{\mathbf{g}_t\}$ is an m.d.s. Since no serial correlation is allowed under this assumption, the asymptotic variance of $\bar{\mathbf{g}}$ is simply $\mathbf{S} = \mathbb{E}(\mathbf{g}_t \mathbf{g}_t')$.
- If there is serial correlation in $\{\mathbf{g}_t\}$, the variance of the limiting distribution of $\sqrt{n}\bar{\mathbf{g}}$ equals the long-run covariance matrix,

$$\mathbf{S} = \sum_{j=-\infty}^{\infty} \mathbf{\Gamma}_j = \mathbf{\Gamma}_0 + \sum_{j=1}^{\infty} (\mathbf{\Gamma}_j + \mathbf{\Gamma}_j'),$$

where $\mathbf{\Gamma}_j$ is the j-th order autocovariance matrix

$$\mathbf{\Gamma}_j = \mathbb{E}(\mathbf{g}_t \mathbf{g}_{t-j}') \quad (j = 0, \pm 1, \pm 2, \ldots).$$

Estimating S When Autocovariances Vanish after Finite Lags

First, the consistent estimators of individual autocovariances are given by

$$\hat{\mathbf{\Gamma}}_j = \frac{1}{n}\sum_{t=j+1}^{n} \hat{\mathbf{g}}_t \hat{\mathbf{g}}'_{t-j} \quad (j = 0, 1, \ldots, n-1)$$

where

$$\hat{\mathbf{g}}_t \equiv \mathbf{x}_t e_t.$$

Given the estimated autocovariances, and if we know that $\mathbf{\Gamma}_j = \mathbf{0}$ for $j > q$ where q is known and finite, then clearly $\mathbf{S}$ can be consistently estimated by

$$\hat{\mathbf{S}} = \hat{\mathbf{\Gamma}}_0 + \sum_{j=1}^{q}(\hat{\mathbf{\Gamma}}_j + \hat{\mathbf{\Gamma}}'_j) = \sum_{j=-q}^{q} \hat{\mathbf{\Gamma}}_j. \tag{14}$$

Using Kernels to Estimate S

The kernel-based estimators can be expressed as a weighted average of estimated autocovariances:

$$\hat{\mathbf{S}} = \sum_{j=-n+1}^{n-1} k\left(\frac{j}{q(n)}\right) \cdot \hat{\mathbf{\Gamma}}_j.$$

Here, the function $k(\cdot)$, which gives weights for autocovariances, is called a kernel and $q(n)$ is called the bandwidth. The estimator (14) is a special kernel-based estimator with $q(n) = q$ and

$$k(x) = \begin{cases} 1 & \text{for } |x| \leq 1, \\ 0 & \text{for } |x| > 1. \end{cases}$$

This kernel is called the truncated kernel. This truncated kernel-based $\hat{\mathbf{S}}$, however, is not guaranteed to be positive semidefinite in finite samples.

[Newey & West, 1987] noted that the kernel-based estimator can be made non-negative definite in finite samples if the kernel is the Bartlett kernel:

$$k(x) = \begin{cases} 1 - |x| & \text{for } |x| \leq 1, \\ 0 & \text{for } |x| > 1. \end{cases}$$

The Bartlett kernel-based estimator of $\mathbf{S}$ is called the Newey-West estimator.

Under certain regularity conditions on random sample, kernel function $k(\cdot)$, and bandwidth $q(n)$, we have

$$\hat{\mathbf{S}} \underset{p}{\to} \mathbf{S},$$

provided $q(n) \to \infty, q(n)/n \to 0$.

Clustered Sampling

In microeconometrics, the observations $\{(y_1, \boldsymbol{x}_1), \ldots, (y_i, \boldsymbol{x}_i), \ldots, (y_n, \boldsymbol{x}_n)\}$ are assumed to be independent and identically distributed. This assumption may be violated if individuals in the sample are connected in some way, for example if they are

- neighbors,
- members of the same village,
- classmates at a school,
- firms within a specific industry.

A currently popular approach which allows for mutual dependence is known as clustered dependence.

In clustering contexts it is convenient to double index the observations as $(y_{ig}, \boldsymbol{x}_{ig})$ where

- $g = 1, \ldots, G$ indexes the cluster
- $i = 1, \ldots, n_g$ indexes the individual within the g^{th} cluster

The total number of observations is $n = \sum_{g=1}^{G} n_g$.

Let $\boldsymbol{y}_g = \left(y_{1g}, \ldots, y_{n_g g}\right)'$ and $\boldsymbol{X}_g = \left(\boldsymbol{x}_{1g}, \ldots, \boldsymbol{x}_{n_g g}\right)'$ denote the $n_g \times 1$ vector of dependent variables and $n_g \times k$ matrix of regressors for the g^{th} cluster. A linear regression model can be written for the individual observations as

$$y_{ig} = \boldsymbol{x}_{ig}'\boldsymbol{\beta} + e_{ig}$$

and using cluster notation as

$$\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{e}_g$$

where $\boldsymbol{e}_g = \left(e_{1g}, \ldots, e_{n_g g}\right)'$ is a $n_g \times 1$ error vector.

The OLS estimator of $\boldsymbol{\beta}$ can be written as

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\boldsymbol{x}_{ig}\boldsymbol{x}_{ig}'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\boldsymbol{x}_{ig}y_{ig}\right)\\ &= \left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{y}_g\right)\\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}\right).\end{aligned}$$

The residuals are $\widehat{e}_{ig} = y_{ig} - \boldsymbol{x}_{ig}'\widehat{\boldsymbol{\beta}}$ in individual level notation and $\widehat{\boldsymbol{e}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g\widehat{\boldsymbol{\beta}}$ in cluster level notation.

The standard clustering assumption is that the clusters are known to the researcher and that the observations are independent across clusters:

Assumption

The clusters $(\boldsymbol{y}_g, \boldsymbol{X}_g)$ are mutually independent across clusters g.

The model is a linear regression under the assumption

$$\mathbb{E}\left[\boldsymbol{e}_g | \boldsymbol{X}_g\right] = 0,$$

or

$$\mathbb{E}\left[e_{ig} | \boldsymbol{X}_g\right] = 0.$$

Note that

$$\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{e}_g\right).$$

We have

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}|\boldsymbol{X}]&=\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\mathbb{E}\left[\boldsymbol{e}_g|\boldsymbol{X}\right]\right)\\&=\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\mathbb{E}\left[\boldsymbol{e}_g|\boldsymbol{X}_g\right]\right)\\&=\mathbf{0}.\end{aligned}$$

Therefore,

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}|\boldsymbol{X}]=\boldsymbol{\beta}.$$

Let

$$\boldsymbol{\Sigma}_g = \mathbb{E}\left[\boldsymbol{e}_g \boldsymbol{e}_g' | \boldsymbol{X}_g\right]$$

denote the $n_g \times n_g$ conditional covariance matrix of the errors within the g^{th} cluster. Since the observations are independent across clusters,

$$\begin{aligned}
\operatorname{Var}\left[\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right) | \boldsymbol{X}\right] &= \sum_{g=1}^{G} \operatorname{Var}\left[\boldsymbol{X}_g' \boldsymbol{e}_g | \boldsymbol{X}_g\right] \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' \mathbb{E}\left[\boldsymbol{e}_g \boldsymbol{e}_g' | \boldsymbol{X}_g\right] \boldsymbol{X}_g \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{\Sigma}_g \boldsymbol{X}_g \equiv \boldsymbol{\Omega}_n.
\end{aligned}$$

It follows that

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \operatorname{Var}[\widehat{\boldsymbol{\beta}} | \boldsymbol{X}] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{\Omega}_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{15}$$

(15) differs from the formula in the independent case due to the correlation between observations within clusters. The magnitude of the difference depends on

- the degree of correlation between observations within clusters
- the number of observations within clusters

Consider a scenario in which $n_g = N$, $\mathbb{E}\left[e_{ig}^2|\boldsymbol{X}_g\right] = \sigma^2$, $\mathbb{E}\left[e_{ig}e_{\ell g}|\boldsymbol{X}_g\right] = \sigma^2\rho$ for $i \neq \ell$, and the regressors $\boldsymbol{x}_{ig}$ do not vary within a cluster. In this case,

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2(1+\rho(N-1)).$$

If $N = 48$ and $\rho = 0.25$, the correct standard errors are a multiple of about three times the conventional formula.

$\boldsymbol{\Omega}_n$ can be estimated by

$$\begin{aligned}\widehat{\boldsymbol{\Omega}}_n &= \sum_{g=1}^{G} \boldsymbol{X}_g' \widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g' \boldsymbol{X}_g \\ &= \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{\ell=1}^{n_g} \boldsymbol{x}_{ig} \boldsymbol{x}_{\ell g}' \widehat{e}_{ig} \widehat{e}_{\ell g} \\ &= \sum_{g=1}^{G} \left(\sum_{i=1}^{n_g} \boldsymbol{x}_{ig} \widehat{e}_{ig} \right) \left(\sum_{\ell=1}^{n_g} \boldsymbol{x}_{\ell g} \widehat{e}_{\ell g} \right)'.\end{aligned}$$

A natural cluster covariance matrix estimator takes the form

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = a_n \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \widehat{\boldsymbol{\Omega}}_n \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1},$$

where a_n is a possible finite-sample adjustment. The Stata cluster command uses

$$a_n = \left(\frac{n-1}{n-k} \right) \left(\frac{G}{G-1} \right).$$

References

Fumio Hayashi. (2000)

Econometrics, Chapter 1, 2, 4 & 6.

Princeton University Press, 2000.

Halbert White. (1980)

A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.

Econometrica **48**: 817–838.

Whitney K. Newey & Kenneth D. West (1987)

A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.

Econometrica **55**: 703-708.